Predicting Academic Course Preference Using Hadoop Inspired Mapreduce

Ms. Namrata Thakre[1], Mr. Hirendra Hajare[2]
[1] Dept of CSE
[2]Assit. Professor, Dept of CSE
[1, 2] Ballarpur Institute of Technology (BIT), Ballarpur.

Abstract- *With the emergence of new technologies, new academic trends introduced into Educational system which results in large data which is unregulated and it is also challenge for students to prefer to those academic courses which are helpful in their industrial training and increases their career prospects. Another challenge is to convert the unregulated data into structured and meaningful information there is need of Data Mining Tools. Hadoop Distributed File System is used to hold large amount of data. The Files are stored in a redundant fashion across multiple machines which ensure their endurance to failure and parallel applications. Knowledge extracted using Map Reduce will be helpful indecision making for students to determine courses chosen for industrial trainings. In this research, we are deriving preferable courses for pursuing training for students based on course combinations. Here, using HDFS, tasks run over Map Reduce and output is obtained after aggregation of results.*

Keywords- Distributed File System, data mining, educational data mining, Hadoop, MapReduce.

I. INTRODUCTION

Data mining is one of the most prominent areas in modern technologies for retrieving meaningful information from huge amount of unstructured and distributed data using parallel processing of data. There is huge advantage to Educational sector of following Data Mining Techniques to analyse data input from students, feedbacks, latest academic trends etc which helps in providing quality education and decision-making approach for students to increase their career prospects and right selection of courses for industrial trainings to fulfil the skill gap pertains between primary education and industry hiring students. Data Mining has great impact in academic systems where education is weighed as primary input for societal progress.

Big data is the emerging field of data mining. It is a term for datasets that are so large or complex that traditional data processing application software is incompetent to deal with them. Big data includes gathering of data for storage and analysis purpose which gain control over operations like searching, sharing, visualization of data, query processing, updating and maintain privacy of information. In Big data, here is extremely large dataset that is analysed computationally to reveal patterns, trends and associations. It deals with unstructured data which may include MS Office files, PDF, Text etc whereas structured data may be the relational data.

Hadoop is one technique of big data and answer to problems related to handling of unstructured and massive data. Hadoop is an open-source programming paradigm which performs parallel processing of applications on clusters. Big Data approach can help colleges, institutions, universities to get a comprehensive aspect about the students. It helps in answering questions related to the learning behaviours, better understanding and curriculum trends, and future course selection for students which helps to create captivating learning experiences for students. The problem of enormously large size of dataset can be solved using Map Reduce Techniques. Map Reduce jobs run over Hadoop Clusters by splitting the big data into small chunks and process the data by running it parallel on distributed clusters.

II. LITERATURE REVIEW

Implementation of Hadoop Operations for Big Data Processing in Educational Institutions

Education plays an important role in maintaining the economic growth of a country. The objective of this paper is to focus on the impact of cloud computing on educational institutions by using latest big data technology to provide quality education. Our educational systems have a large amount of data. Big Data is defined as massive sets of data that is so large or so complex that it is very difficult to process by using conventional applications and software technologies. This has resulted in the penetration of Big Data technologies and tools into education, to process the large amount of data involved. In this paper we discuss what Cloud and Hadoop is, and its types, operations and services offered. Hence it has an advantage which will surely help the students when used in an appropriate way.

Predicting Student Performance Using Map Reduce Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on Student datasets using multiple classifiers and feature selection techniques. Many of them show good classification accuracy. The existing work proposes to apply data mining techniques to predict Students dropout and failure. But this work doesn't support the huge amount of data. It also takes more time to complete the classification process. So, the time complexity is high. To improve the accuracy and reduce the time complexity, the Map Reduce concept is introduced. In this work, the deadline constraint is also introduced. Based on this, an extensional Map Reduce Task Scheduling algorithm for Deadline constraints (MTSD) is proposed. It allows user to specify a job's (classification process in data mining) deadline and tries to make the job to be finished before the deadline. Finally, the proposed system has higher classification accuracy even in the big data and it also reduced the time complexity.

Map Reduce as a Programming Model for Association Rules Algorithm on Hadoop as association rules widely used, it needs to study many problems, one of which is the generally larger and multi-dimensional datasets, and the rapid growth of the amount of data. Single processor's memory and CPU resources are very limited, which makes the algorithm performance inefficient. Recently the development of network and distributed technology makes cloud computing a reality in the implementation of association rules algorithm. In this paper we describe the improved Apriori algorithm based on Map Reduce mode, which can handle massive datasets with a large number of nodes on Hadoop platform.

Apriori-Map/Reduce Algorithm Map/Reduce algorithm has received highlights as cloud computing services with Hadoop frame works were provided. Thus, there have been many approaches to convert many sequential algorithms to the corresponding Map/Reduce algorithms. The paper presents Map/Reduce algorithm of the legacy Apriori algorithm that has been popular to collect the item sets frequently occurred in order to compose Association Rule in Data Mining. Theoretically, it shows that the proposed algorithm provides high performance computing depending on the number of Map and Reduce nodes.

III. PROPOSED SYSTEM

Using Map Reduce, the application can be scaled to run over multiple machines in a cluster and for that Hadoop cluster is used. The Map Reduce Framework consists of Map and Reduce Functions with single Resource Manager which acts as a master and one Node manager which acts as slave per cluster node. The input dataset is fed into the mapper and after passing through shuffle phase, reducer displays the output after aggregating the tuples obtained from mapper and are in the form of <key, value> pair.

IV. METHODOLOGY

Dataset for Course Selection

Table I shows the list of course combinations taken by students for their industrial trainings

Table I. Dataset for Course Selection

Course Combination	Preferable Course
Java,J2EE	Android
HTML,Javascript	PHP
C,C++ Asp.Net	Asp.Net
J2EE,Java	Android
C,C++	Java

Here each row is considered as transaction, each comprising a combination of variables or item sets. The pattern of student's choice for industrial training course combinations is predicted after processing through MAP Reduce Hadoop Data Mining Technique shown in figure.

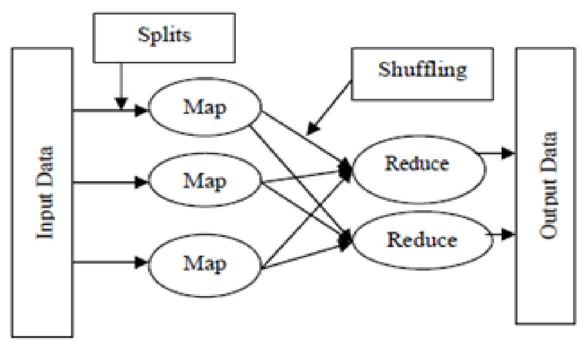

Fig 1: MapReduce Organization Chart

The input dataset collected from students is shown in Table I is stored in the HDFS for MapReduce. The input data is then split into various clusters and provide it to the mapper that maps data to the output. The output from the mapper is represented as <key, value> pair. The output obtained from the mapper are then combined together in the combiner and then sent to the reducer.

Here, for organizing the work, Hadoop divides the task into Map and Reduce Tasks. The components of Hadoop Distributed File System are discussed below. The Map Reduce program transforms lists of input data elements into list of output data elements and it will be done using twice by Map and Reduce. The cluster running applications and name node information in the form of Web Interfaces using Hadoop.

The Organisation Structure of Map Reduce Framework is shown in Fig.1which represents that input data obtained from Table 1 first splits and then mapped followed with shuffling. The unstructured data after shuffling filtered to obtain output which is also called "Reduce Phase".

A. Mapper

In Map function, individual jobs transform records into intermediate records. There is multiplicity in the input pair which map to zero or one to many output pairs.

B. Reducer

In Reducer Function, the set of intermediate value share key of smaller set of values which reduces the overhead of the system. In reducer, output is obtained after merging.

C. Name Node

The Name node is the main feature of Hadoop Distributed File System. The name node stores all the metadata for the file system, using yarn command, the Fig.3 and 4 shows the node manager and resource manager running concurrently. The Name node uses RAM space.

D. Data Node

Data node stores the actual data in HDFS. Here, data node is known as Slave and Name node is known as Master. There is Master-Slave communication between Name node and Data node. When data node starts, it communicates to the Namenode along with blocks list managed by it. The Datanode uses Hard disk space.

E. Resource Manager

The progress of the jobs running in cluster can be viewed through Resource Manager Web interface shown in Fig.6. Along with it, the status of the scheduler can also be viewed. The resource manager determines all the available cluster resources and help in managing the distributed applications. It works with Node Manager and Application Master.

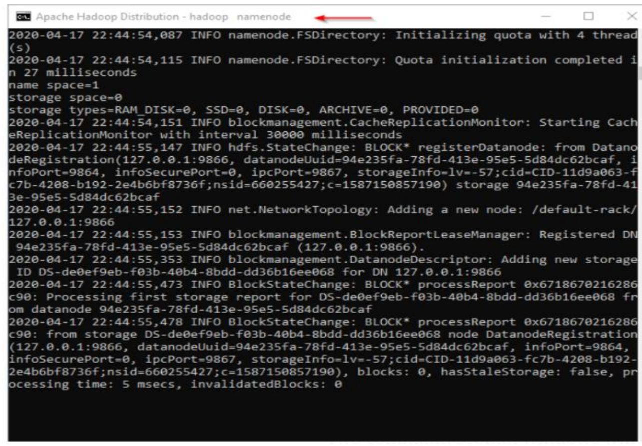

Fig.2 Hadoop Name Node

F. Job Tracker

Job tracker plays an important role for complete execution of the tasks submitted. Job tracker resides on name node when only one task to be accomplished and for multiple tasks, job tracker resides on Data node. The overall progress of each job is tracked through Job Tracker.

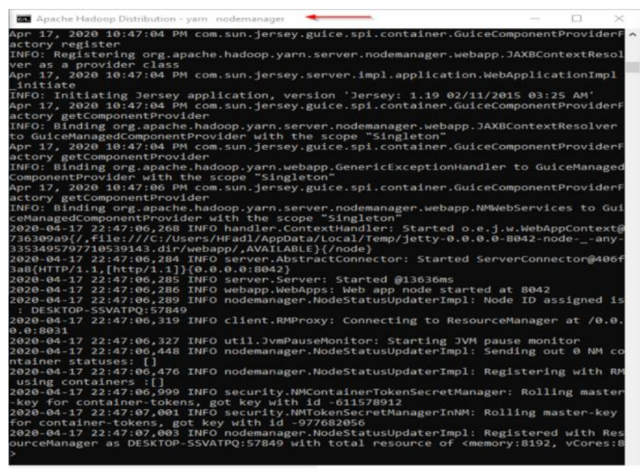

Fig.3 Hadoop Node Manager

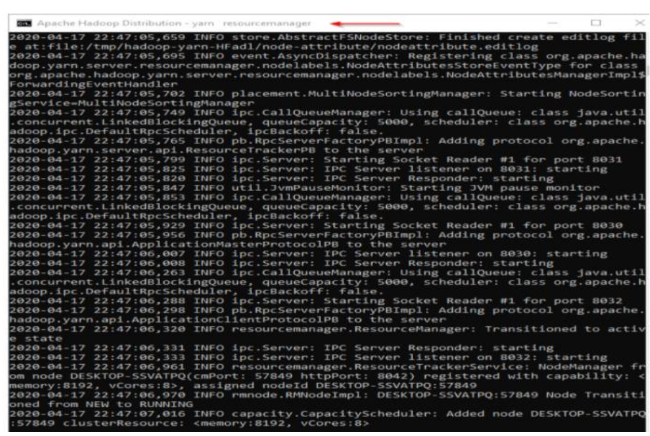

Fig.4 Yarn Resource Manager

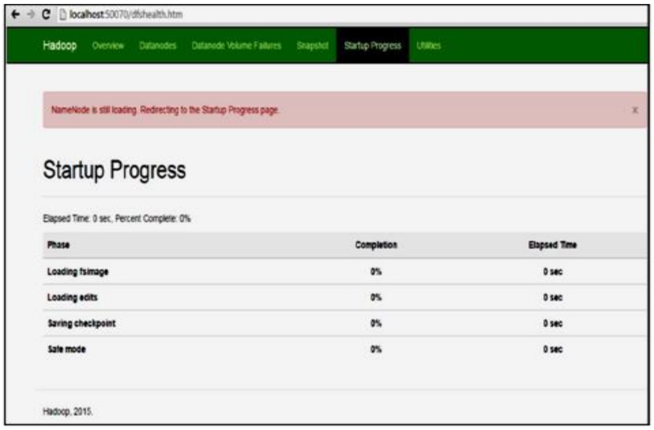

Fig.5 Web Interface of Name Node

The features and attributes shown in Figure 5 and 6 are tabulated in Table II. In Table Cluster Metrics and Scheduler Metrics are shown with Memory Consumption and Hadoop Distributed File System Health.

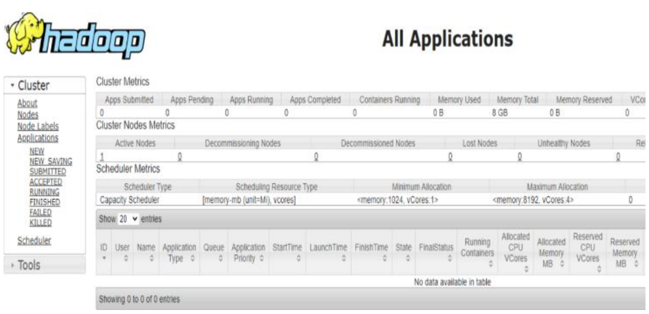

Fig.6 Web Interface of Resource Manager displaying Cluster Information

V. RESULT & FUTURE SCOPE

After processing the input data through Map Reduce in Hadoop, display the desired output. The input dataset splits and after mapping, process of shuffling is performed and the output of Mapper becomes input to Reducer function and after this final output is obtained. The result shows that maximum of students have shown key interest towards “C, C++ and Java Courses” which not only helps management as well Faculty members for more focus on above mentioned courses.

VI. CONCLUSION

The Map Reduce approach is used for running jobs over HDFS. Using Map Reduce, the application can be scaled to run over multiple machines in a cluster and for that Hadoop cluster is used. The Map Reduce Framework consists of Map and Reduce Functions with single Resource Manager which acts as a master and one Node manager which acts as slave per cluster node. The input dataset is fed into the mapper and after passing through shuffle phase, reducer displays the output after aggregating the tuples obtained from mapper and are in the form of <key, value> pair.

REFERENCES

[1] Saptarshi Ray, “Big Data in Education”, Gravity, the Great Lakes Magazine, pp. 8-10, 2013.

[2] Jongwook Woo, "Apriori-Map/Reduce Algorithm." Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2012.

[3] Katrina Sin,LoganathanMuthu,"Application of Big Data in Education Data Mining and Learning Analytics – A Literature Review" ,ICTACT Journal on Soft Computing,ISSN: 2229-6956 (online), Vol 5, Issue 4,July 2015.

[4] Shriram Raghunathan and Abtar Kaur, “Assessment of online interaction pattern using the Q-4R framework”, The International Lifelong Learning Conference, 2011.

[5] B.Manjulatha,Ambica Venna, K.Soumya, "Implementation of Hadoop Operations for Big Data Processing in Educational Institutions",International Journal of Innovative Research in Computer and Communication Engineering,ISSN(Online) : 2320-9801,Vol. 4, Issue 4, April 2016.

[6] N.Tajunisha,M.Anjali, "Predicting Student Performance Using MapReduce”,IJECS,Vol.4,Issue 1, Jan 2015, p. 9971-9976.

[7] Shankar M.Patil,Praveen Kumar, “Data Mining Model for Effective Data Analysis of Higher Education Students Using MapReduce”,IJERMT,ISSN:2278-9359,Vol.6,Issue4,April2017.

[8] Madhavi Vaidya,"Parallel Processing of cluster by Map Reduce",IJDPS,Vol.3,No.1,2012.

[9] HarshawardhanS.Bhosale,DevendraP.Gadekar,"A Review Paper on Big Data and Hadoop",IJSRP,ISSN:2250-3153,Vol 4,Issue 10,Oct 2014.